



Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer

Dr Dimitris Polychronopoulos (Oncology R&D)

Dr Anna Gogleva (R&D IT)

17th July 2022, ICML, Baltimore, USA



Global, science-led, patient-focused biopharmaceutical company



Science and innovation-led



Therapy areas of focus:
Oncology;
Cardiovascular, Renal & Metabolism;
Respiratory & Immunology;
Rare Disease



Diversified portfolio with broad coverage across primary care, specialty care and rare diseases



Commitment to people and society



Global strength, with balanced presence across regions



Strategic R&D sites close to global bioscience clusters



Focus on main therapy areas and key platforms



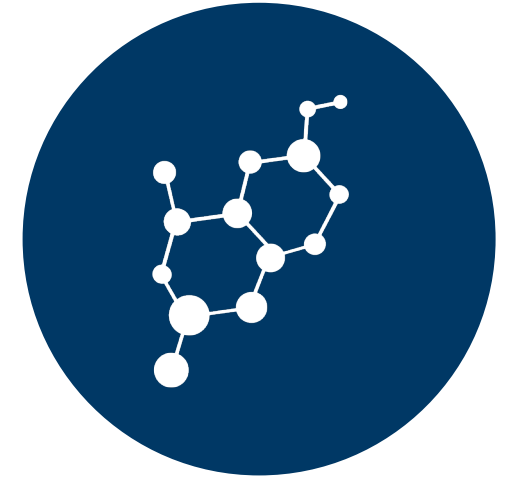
Oncology



Cardiovascular,
Renal &
Metabolism



Respiratory &
Immunology



Rare Disease



Small
molecules

Biologics

Protein
engineering

Complement
inhibition

Other
emerging drug
platforms

Diagnostics

Devices



Oncology

We are leading a revolution in oncology to redefine cancer care and Data Science plays a critical part



Our clinical strategy is designed to help transform survival



With our portfolio and pipeline we strive to revolutionise cancer care



Catalysing changes in the practice of medicine to transform the patient experience



We are driven by our passion, our people and a culture of innovation



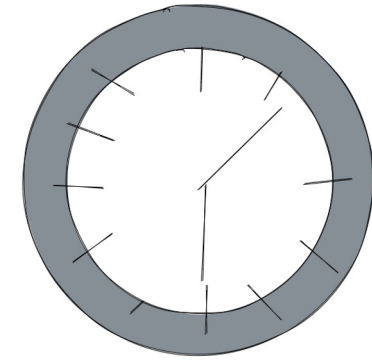
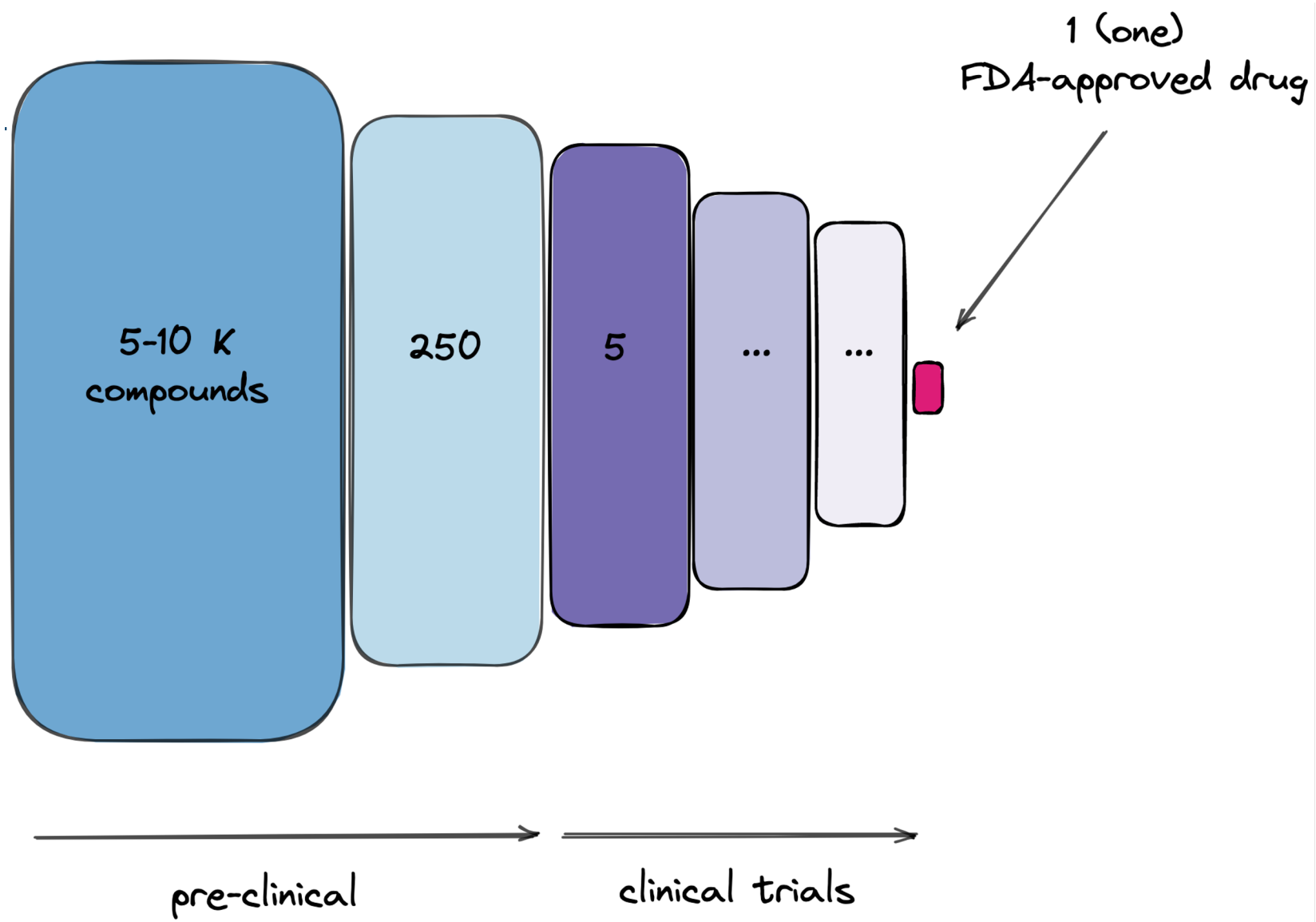
Life-cycle of a medicine

entire life-cycle of a medicine:

- research and development
- manufacturing and supply
- global commercialisation



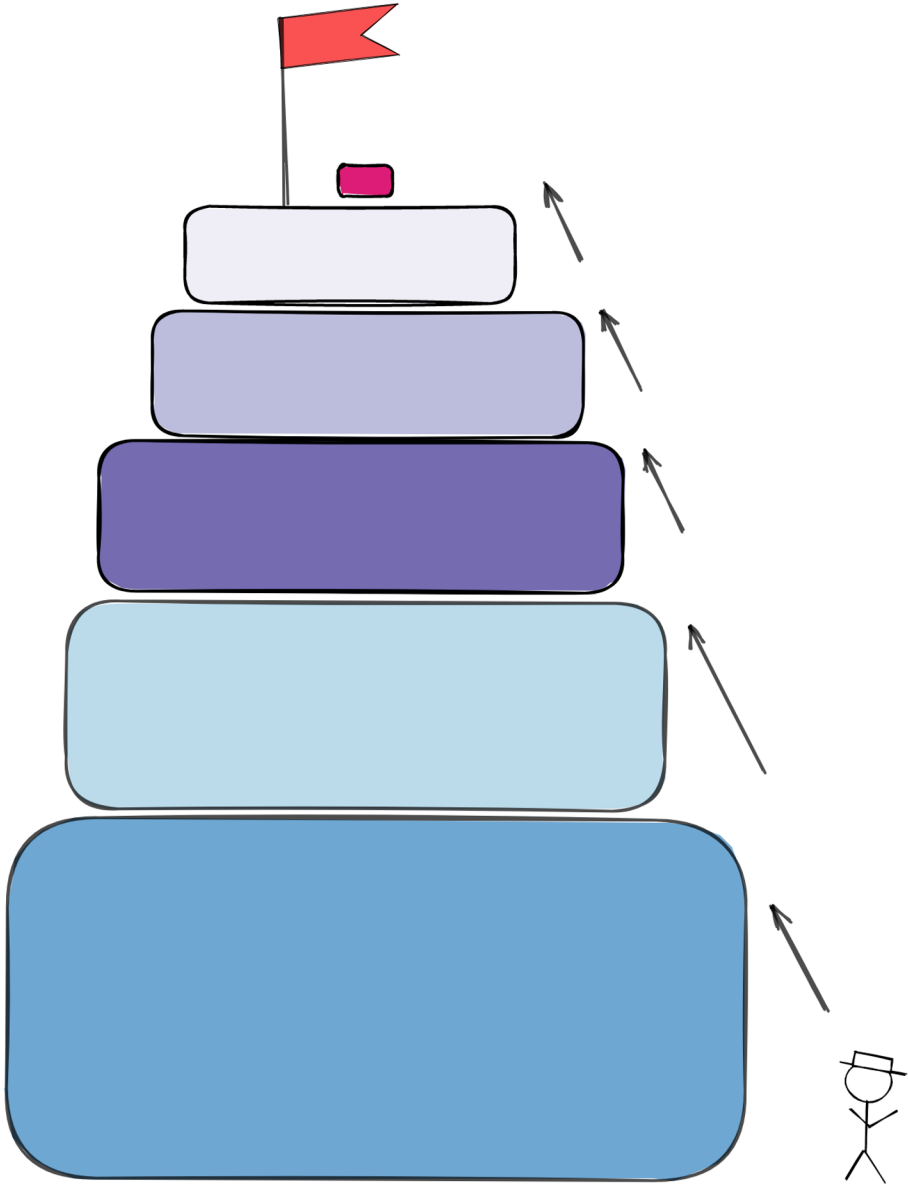
One needs to fail a lot to discover a working drug



6 - 10
years



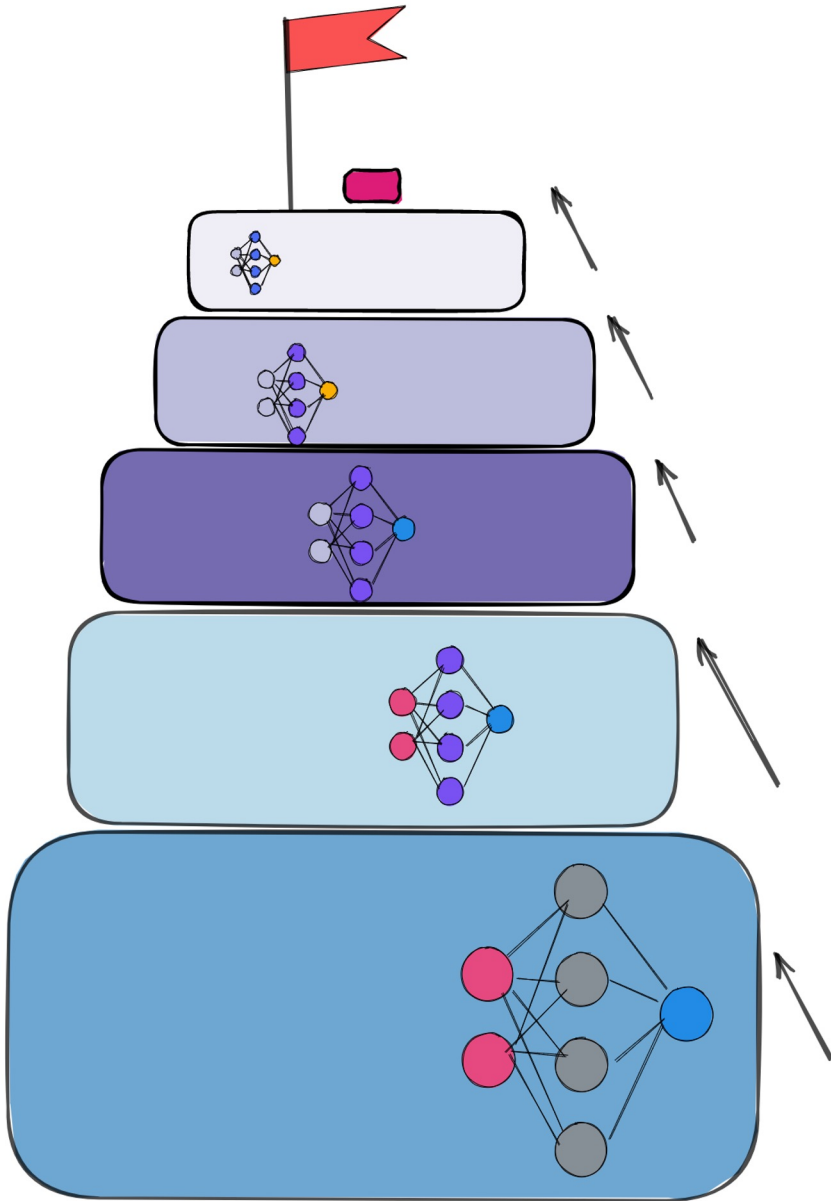
It is a tall mountain to climb



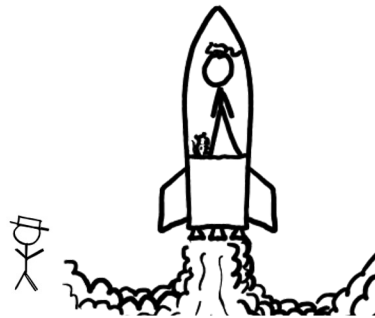
- How to develop new efficient treatments faster?
- How to make better decisions in the process?



It is a tall mountain to climb

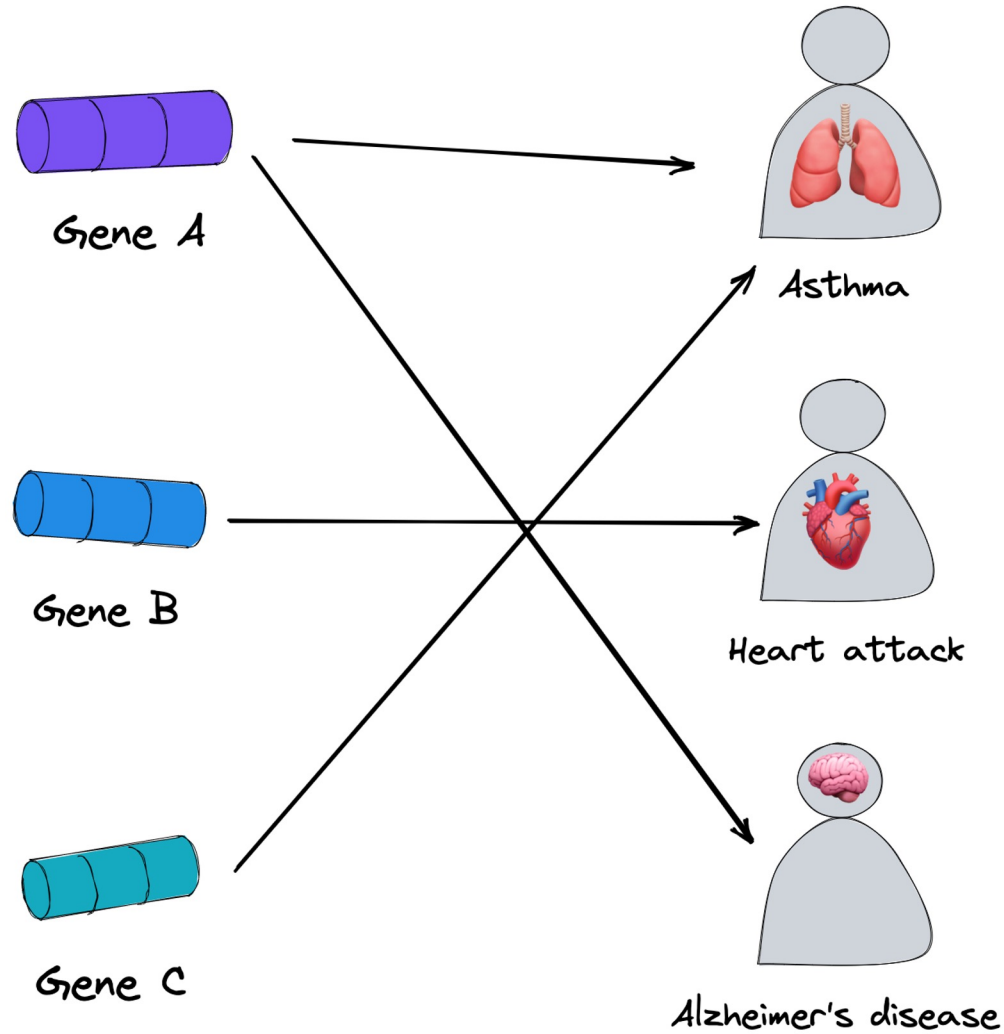


- How to develop new efficient treatments faster?
- How to make better decisions in the process?
- Recommendation systems can help in multiple places

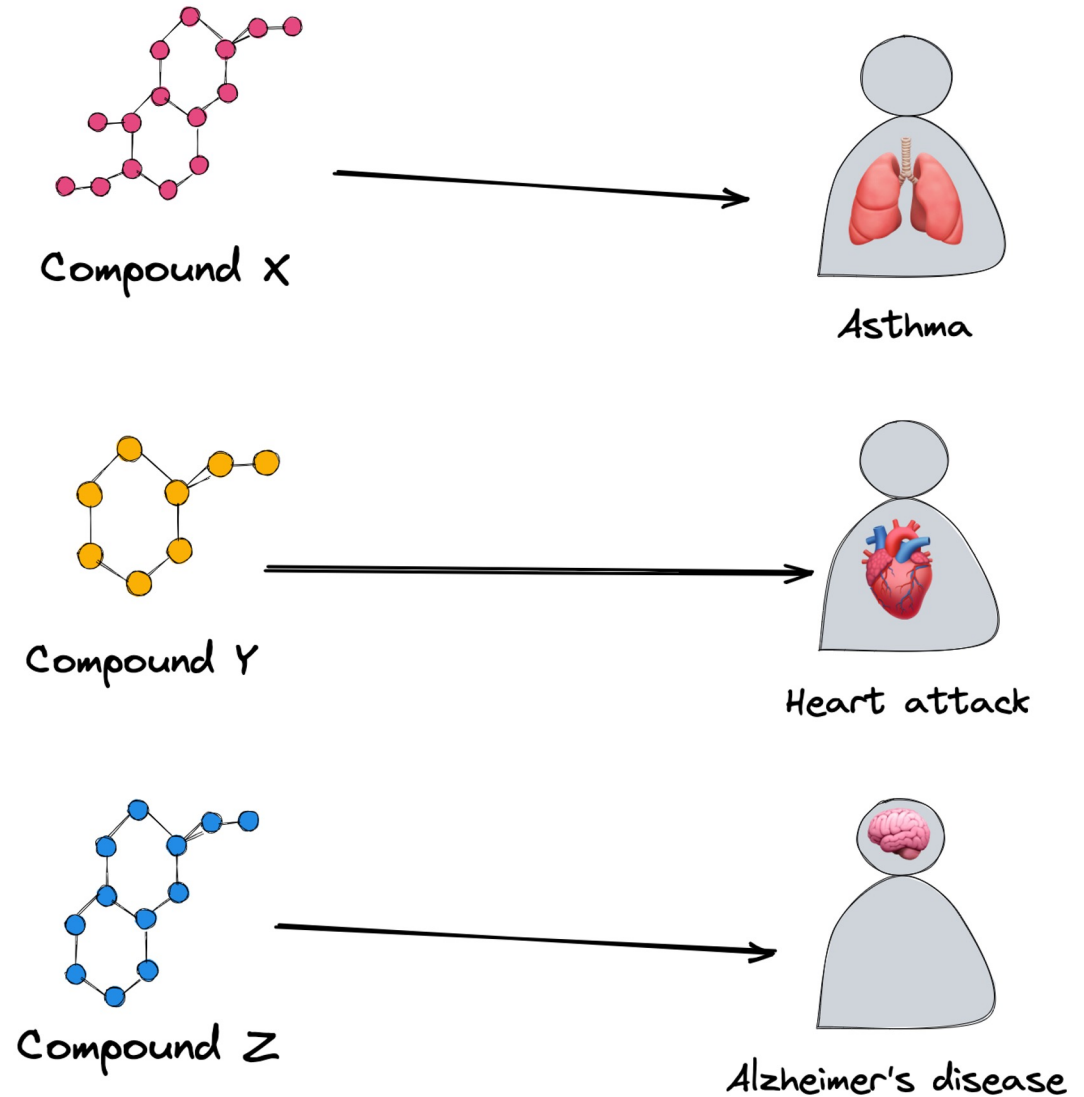


Recommendation problems in drug discovery

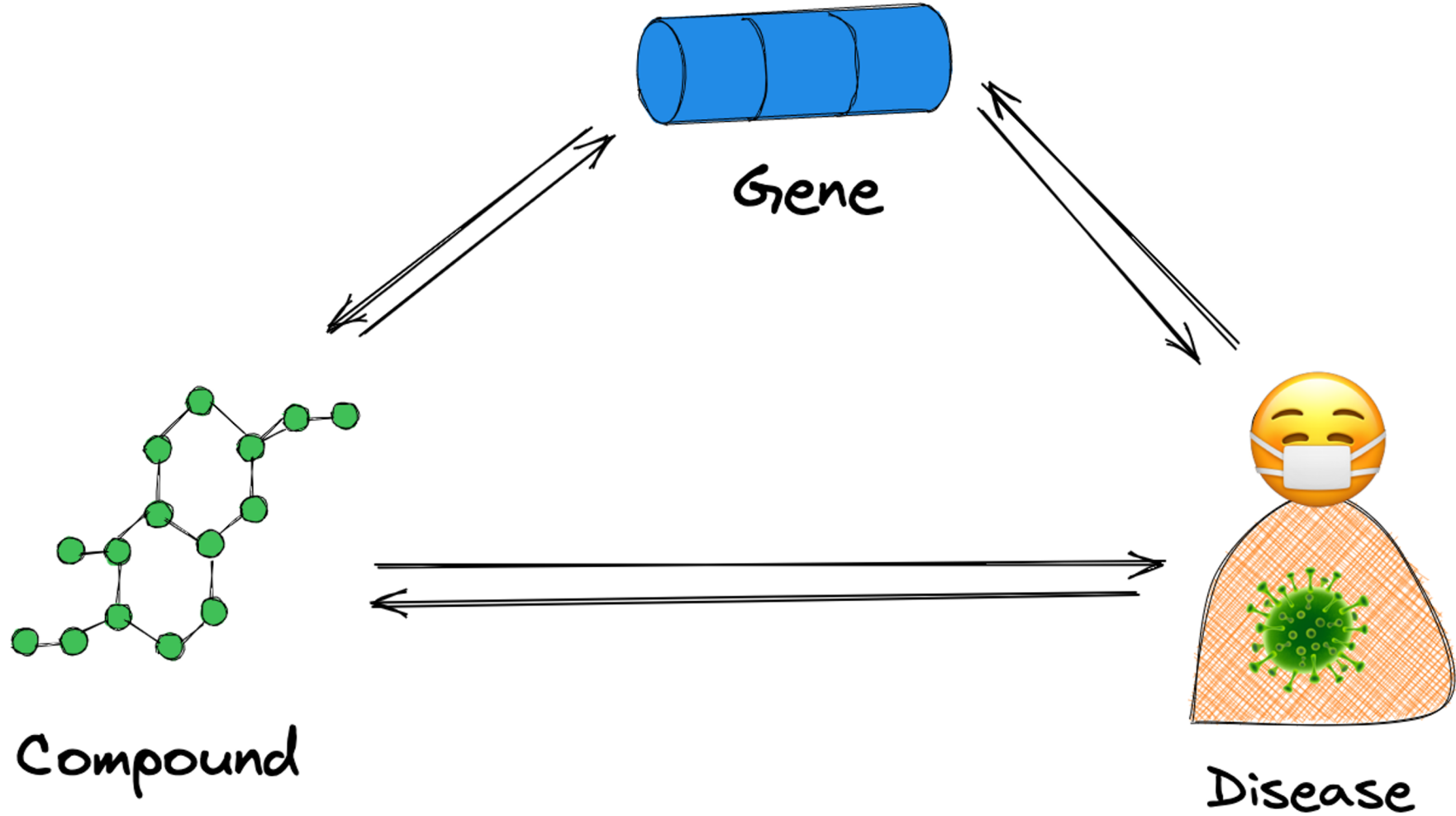
find a gene causing a disease



match a drug with a disease



Drugs, genes, diseases

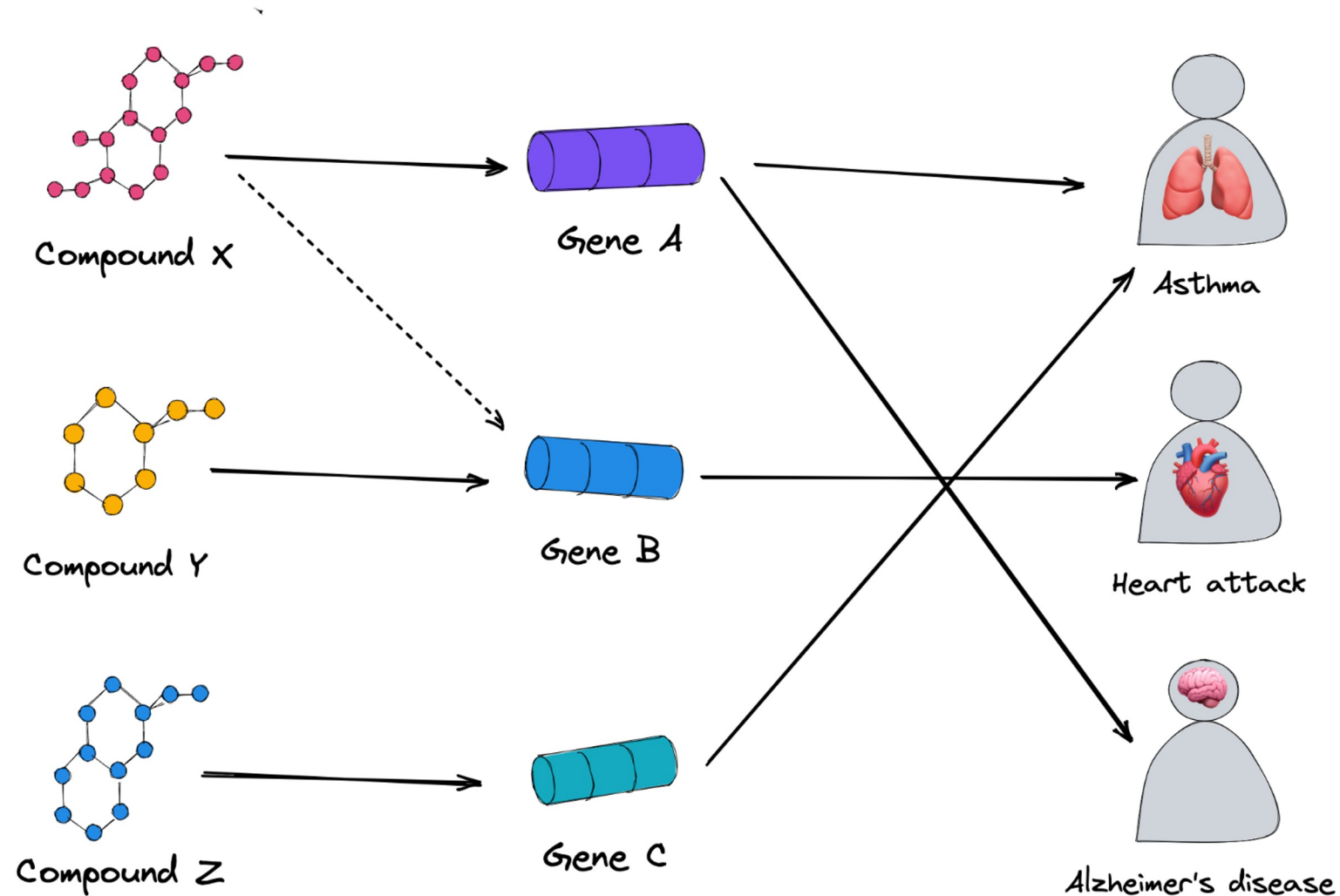


It gets complex very fast

Millions of compounds
Billions possible theoretically

25-30 K genes,
80 K functional elements

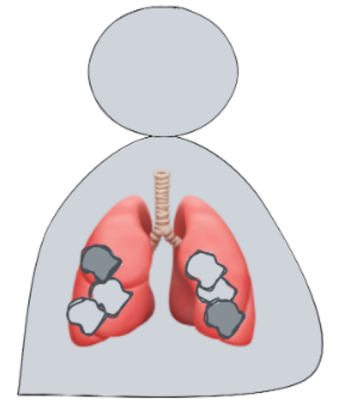
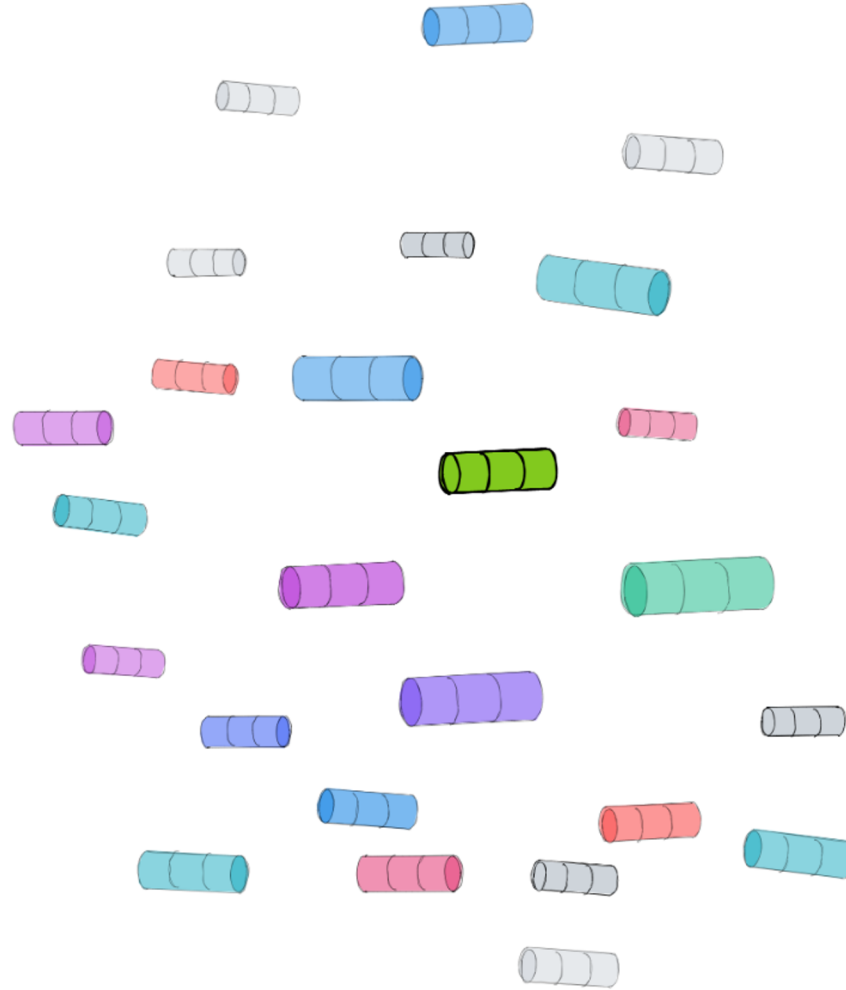
~10 K diseases



It is rarely just a single gene

● 25-30K human genes

● everything interacts with everything,
each gene is a suspect

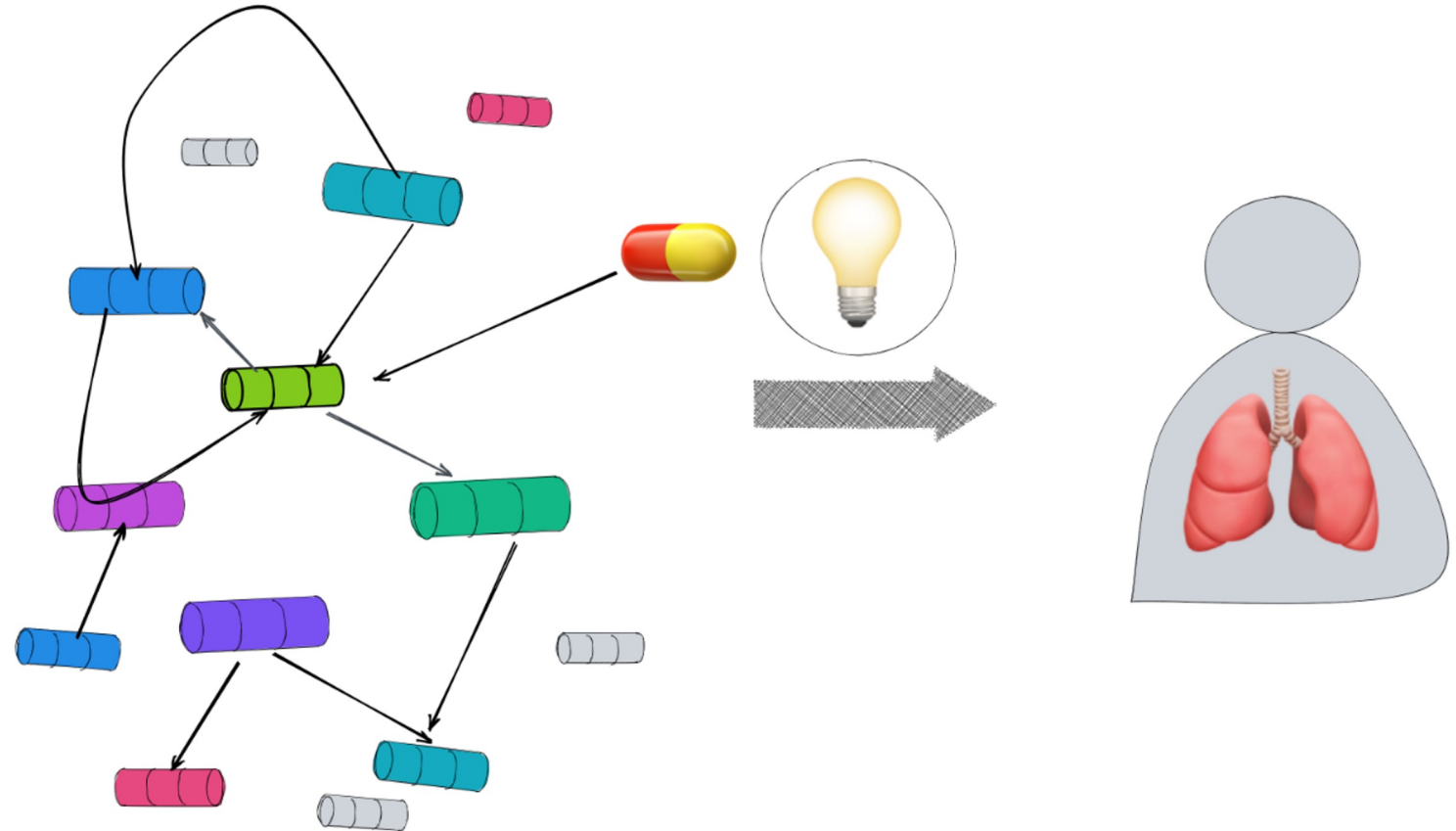


a disease



Find a molecular network behind a disease

- 1 disease ~ a molecular process gone awry
- 2 find the key molecular process
- 3 re-route it safely



Biomedical knowledge is spread across multiple resources

depmap

BioGrid
Connecting health information

uberon

HGNC

PubMed

Cellosaurus

e!Ensembl

MedDRA

OMIM
Human Genetics Knowledge
for the World

ChEMBL

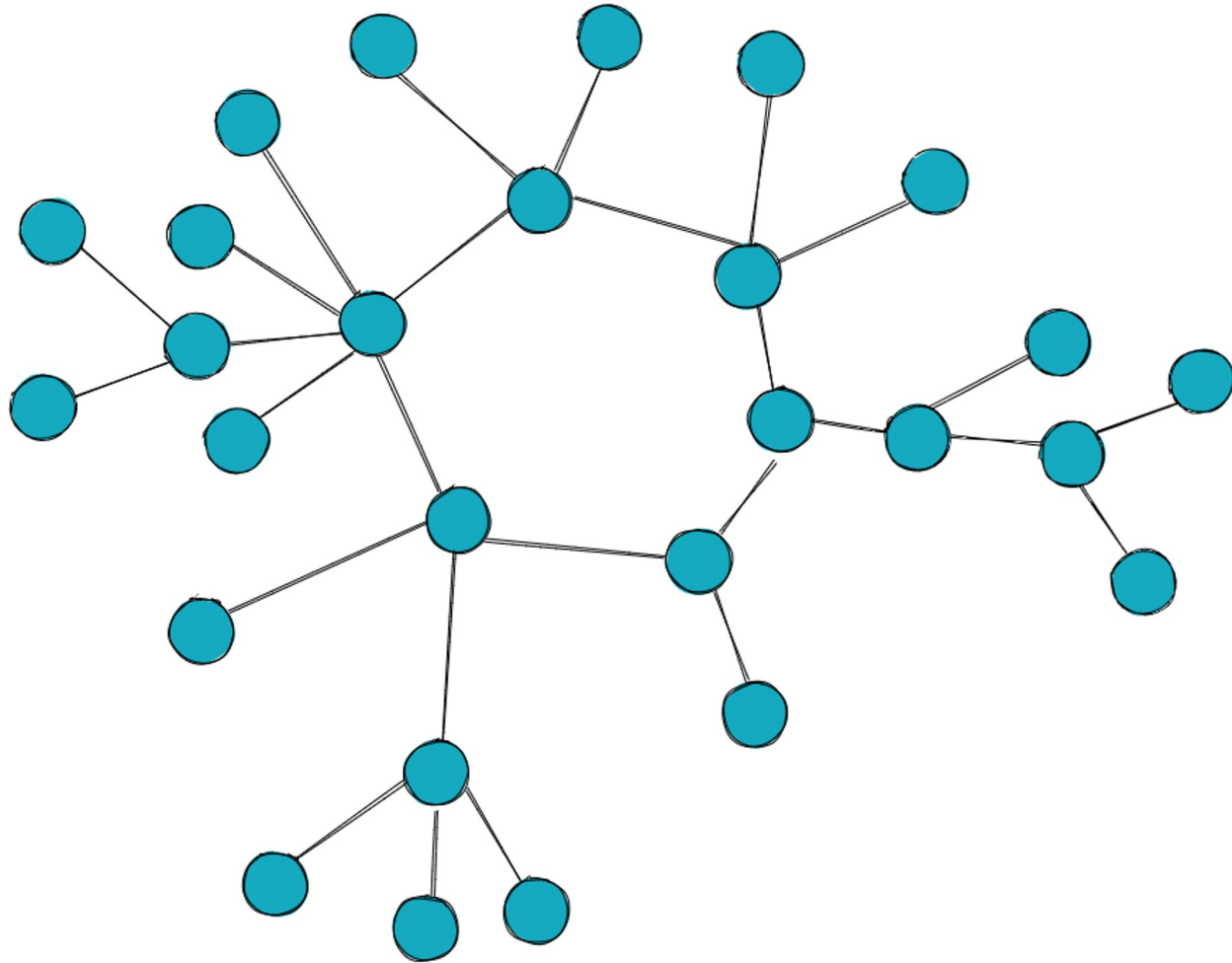
DisGeNET

mondo
THE WORLD'S DISEASE CONCEPTS, UNIFIED

EBI
Expression Atlas

Global Online Structure Activity Relationship Database

Graph makes things simpler



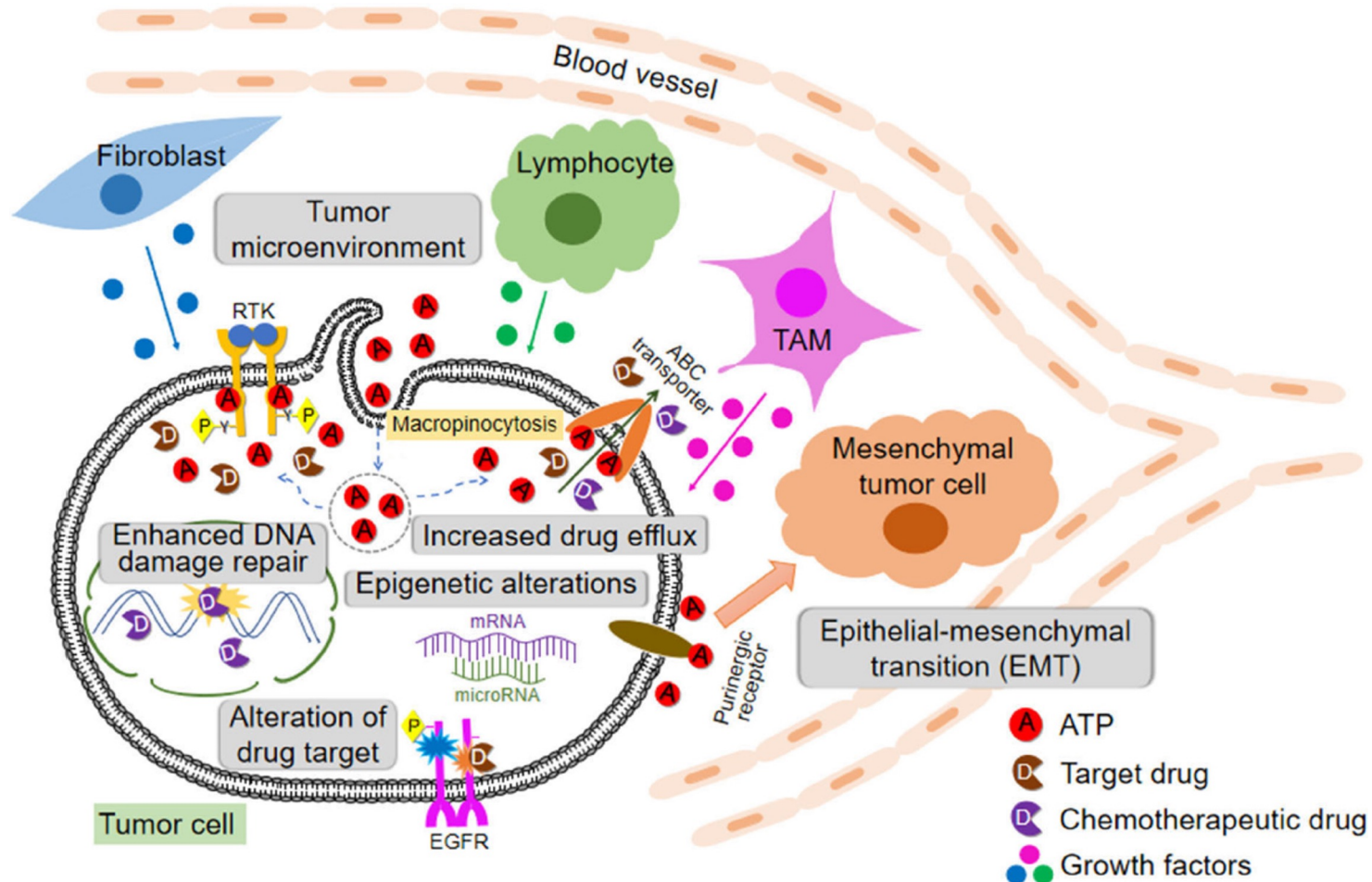
- Biomedical information often comes in forms of **networks** and **hierarchies**
- Graph is a convenient way to organize it
- BIKG (our internal knowledge graph): **60+** data sources including - omics and data extracted from the literature
- **11 M nodes, 1 B edges**
- Use graph as a source of context and features for recommenders



Early success story: graph-based recommendations



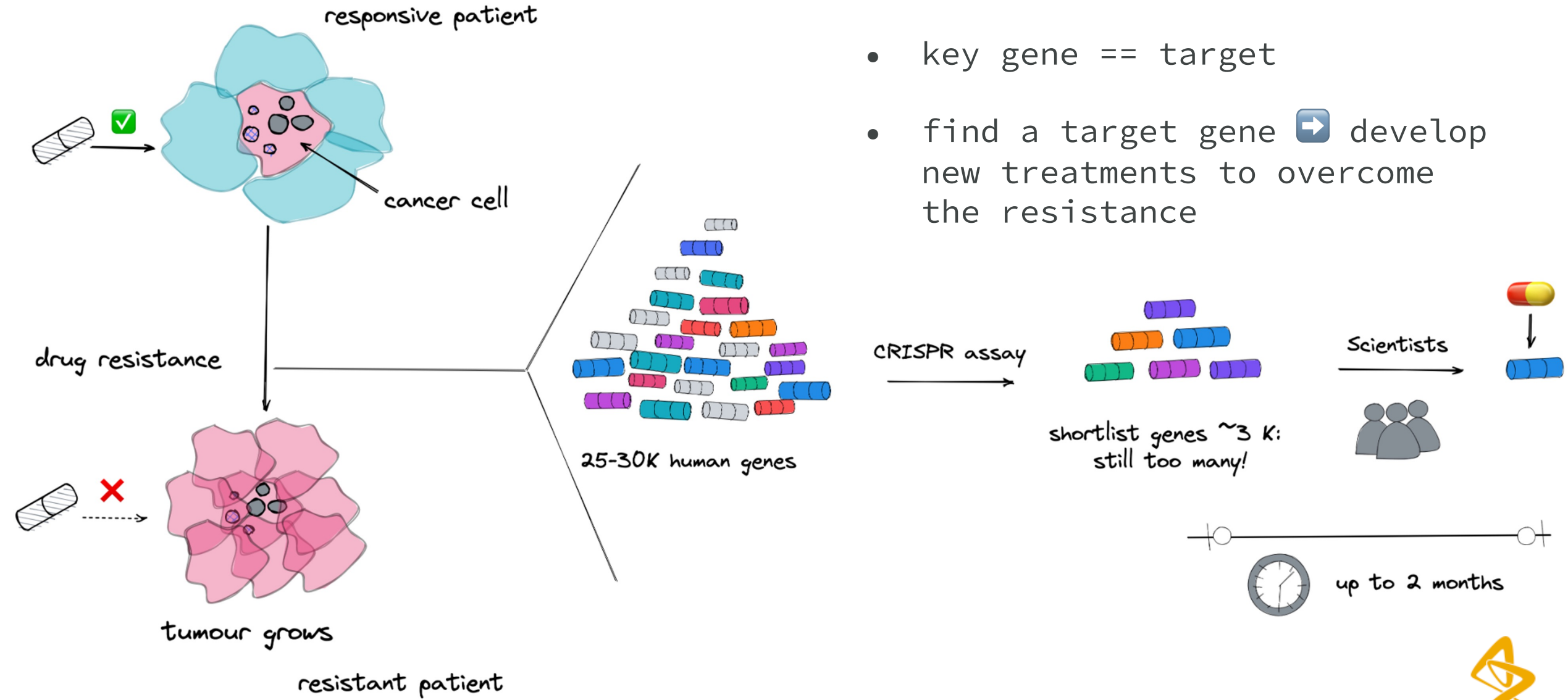
Applied recommendation problem: contextualize experimental data



- Drug resistance in lung cancer
- Occurs in a sub-population of patients
- Resistance landscape is complex



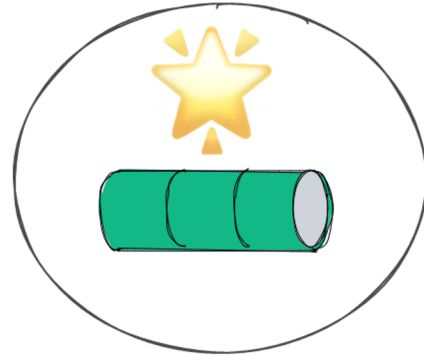
How to help scientist find key genes faster?



- key gene == target
- find a target gene → develop new treatments to overcome the resistance



An ideal target

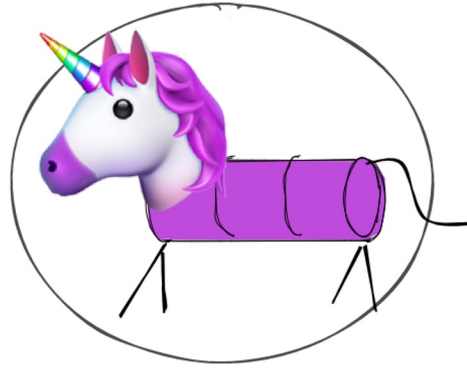


- Expression
- Pathway/complex enrichment
- Effect size
- Druggability
- Mode of action
- Translation in models
- Internal assets
- Bench validation
- Consistency in assays
- Clinical relevance
- Literature support
- Novelty

...



An ideal target does not exist

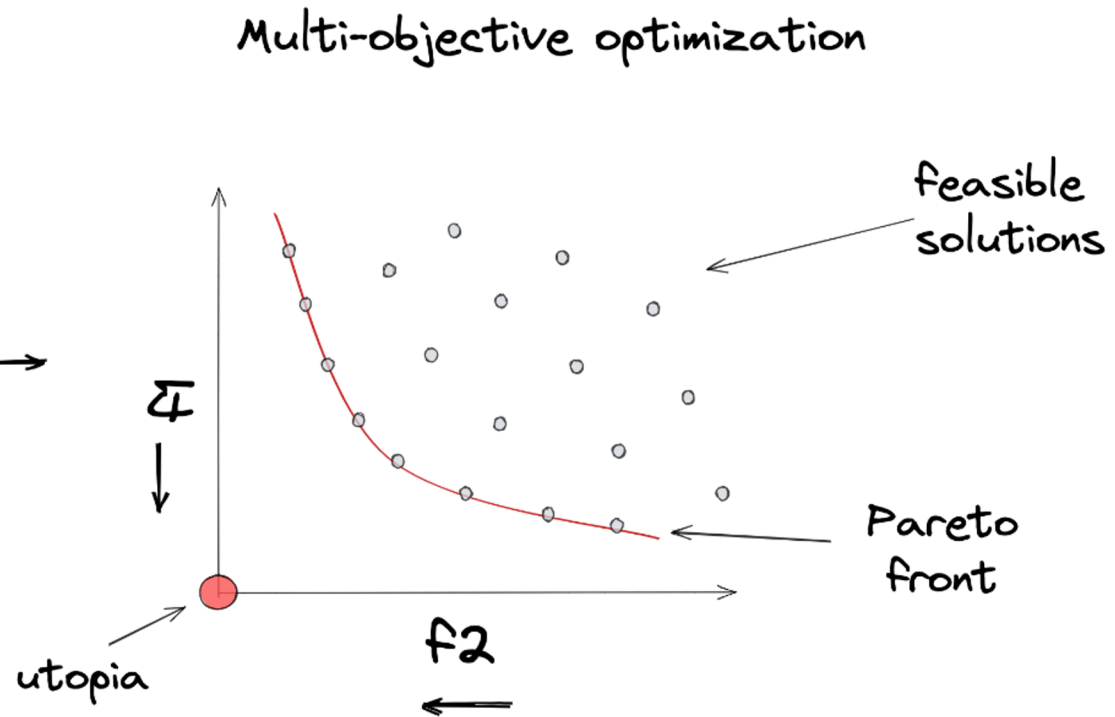
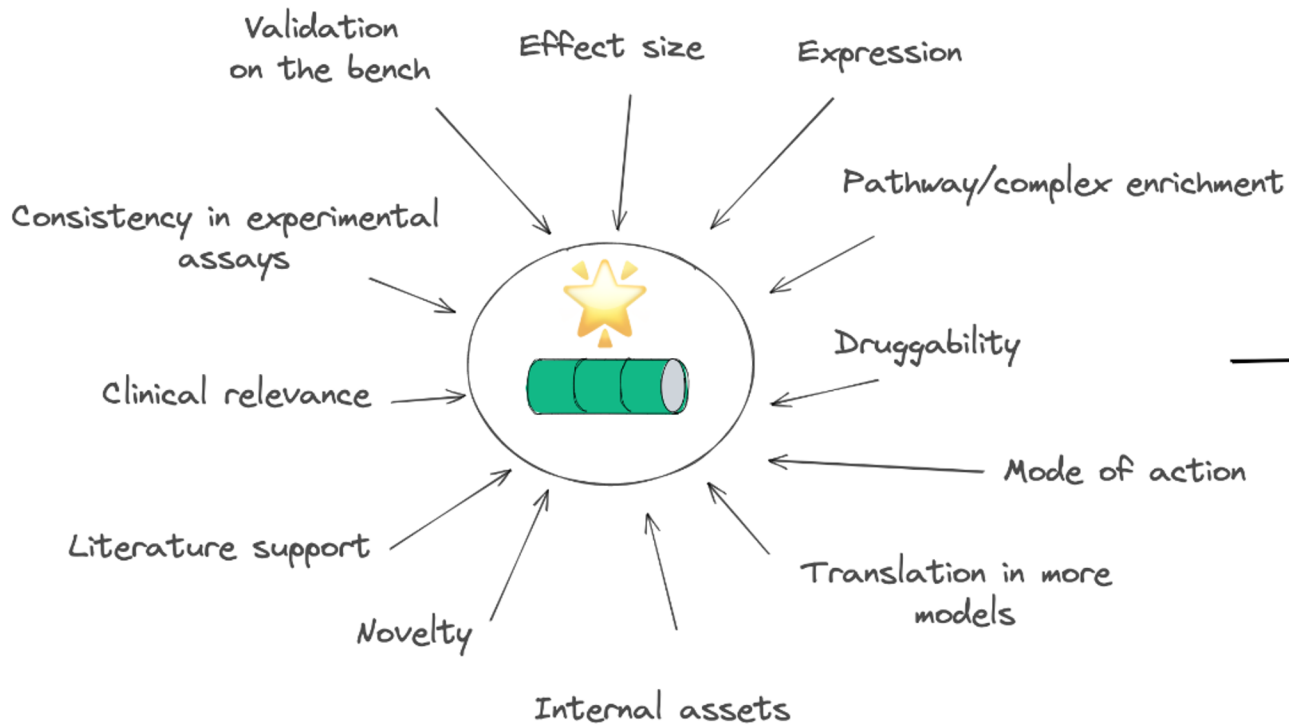


- Expression
- Pathway/complex enrichment
- Effect size
- Druggability
- Mode of action
- Translation in models
- Internal assets
- Bench validation
- Consistency in assays
- Clinical relevance
- Literature support
- Novelty

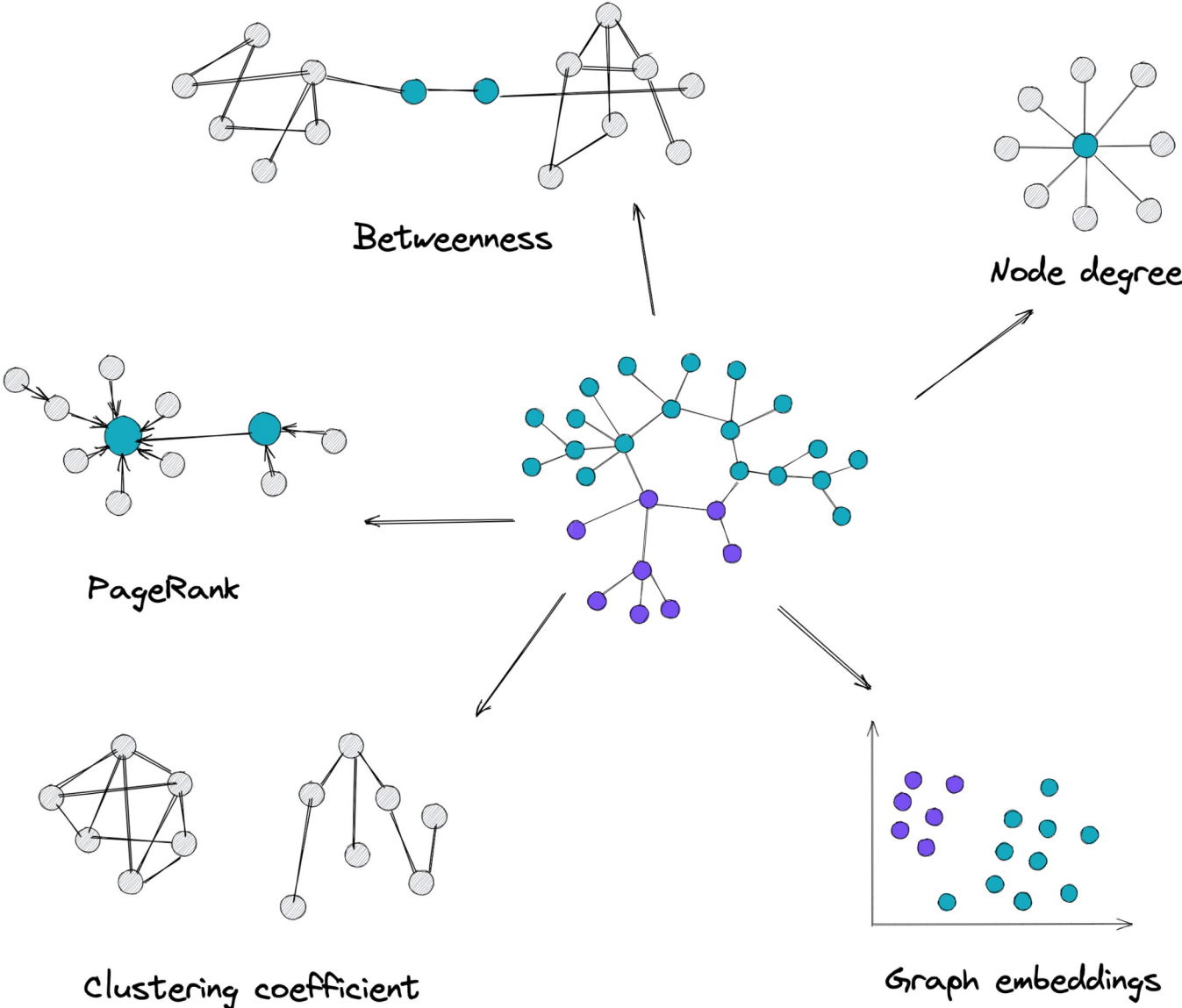
...



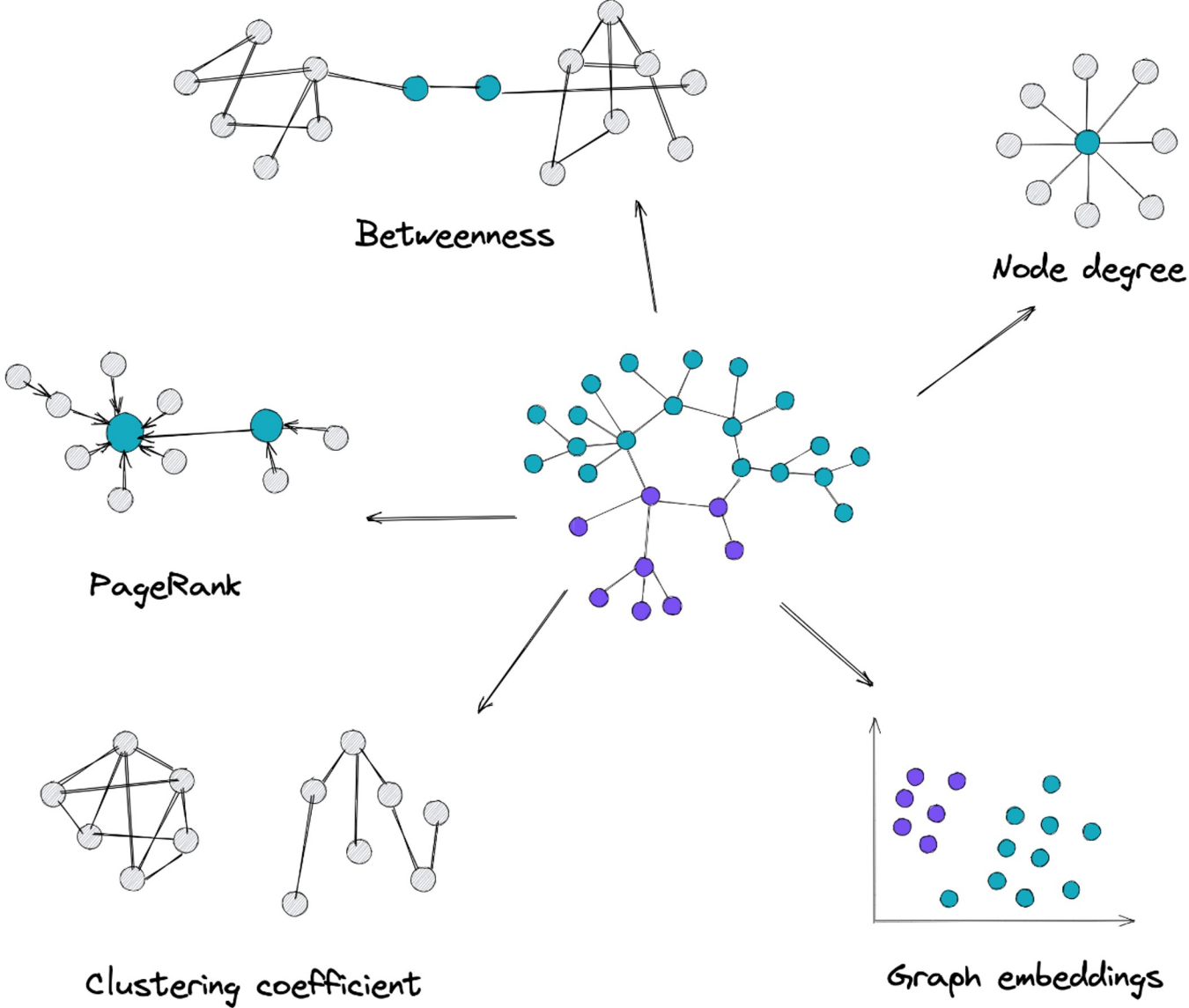
Target selection as an optimization problem



Hybrid feature set: source features from the graph



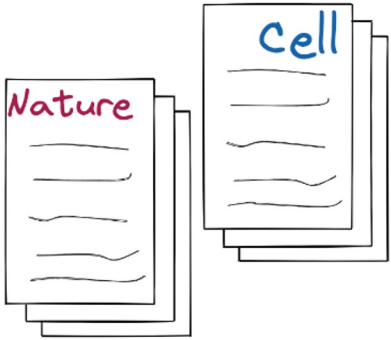
Hybrid feature set: combine with clinical features



Clinical features



Literature support



Druggability

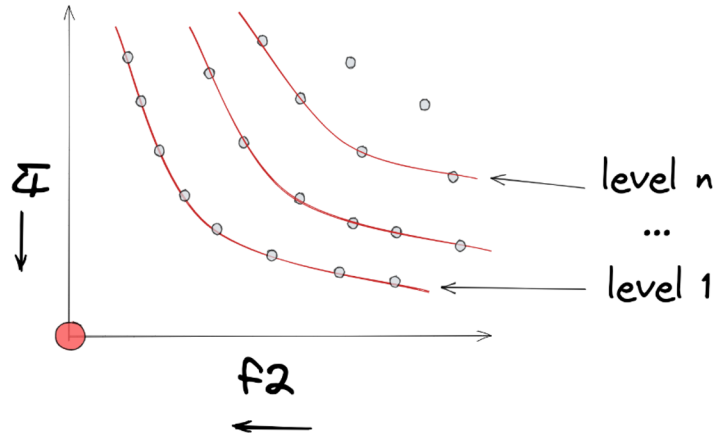


Pre-clinical experimental assays

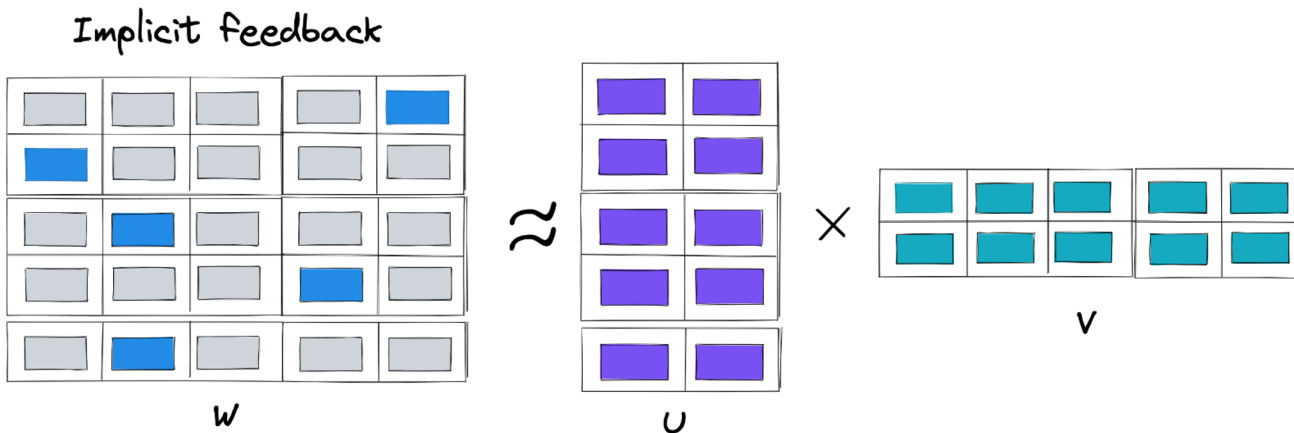


Approaches

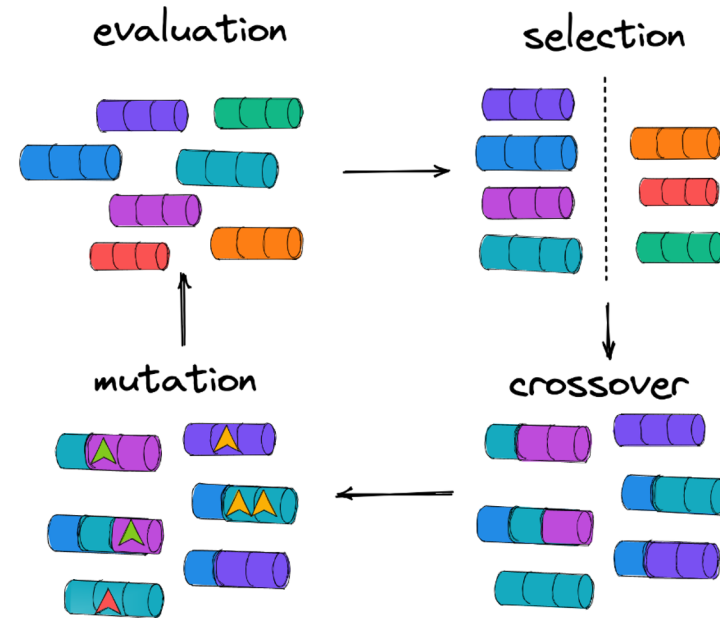
① Compute exact Pareto front



③ Matrix factorization



② Evolutionary algorithms



SkywalkR, interactive interface

- select a subset of objectives
- set optimization directions
- explore trade-offs

The screenshot displays the SkywalkR web interface. On the left is a sidebar with navigation options: Essentiality, Tractability, BIKG, literature support, BIKG, graph-derived, Consistency, Clinical relevance, and Preclinical evidence. Below these are three enrichment score sliders for 'Tesla, enrichment score, RESPONDERS vs RESISTANT', 'Flaura, enrichment score, RESPONDERS vs RESISTANT', and 'Orchard, enrichment score, #ORCHARD vs #FLAURA pre-treatment'. At the bottom of the sidebar are 'rank!' and 'reset' buttons.

The main content area shows a search for 'How it works' and a 'Result' tab. It includes a 'Sort top genes by' dropdown set to 'full_screen' and a 'Column visibility' section showing 'Show 10 entries'. Below this is a table of gene results:

gene	trct_sm	nlp_egfr	nlp_nslc	full_screen	KO_osi	KO_gefi	KO_all	A_osi	A_gefi	A_all	tesla_ES
WWTR1	6	0	0	8	0	0	0	5	3	80.00	
KCTD5	0	0	0	8	5	3	8	0	0	0	0.00
NF1	6	0	0	7	4	3	7	0	0	0	0.33
FOSL1	0	0	0	6	0	0	0	4	2	6	0.00
MET	9	8	66	6	0	0	0	4	2	6	0.67
PTEN	6	33	0	6	4	2					
NF2	6	0	0	6	4	2					
CSF1R	9	0	0	5	0	0					
TSC2	0	0	0	5	3	2					
KFAP1	9	7	0	5	2	2					

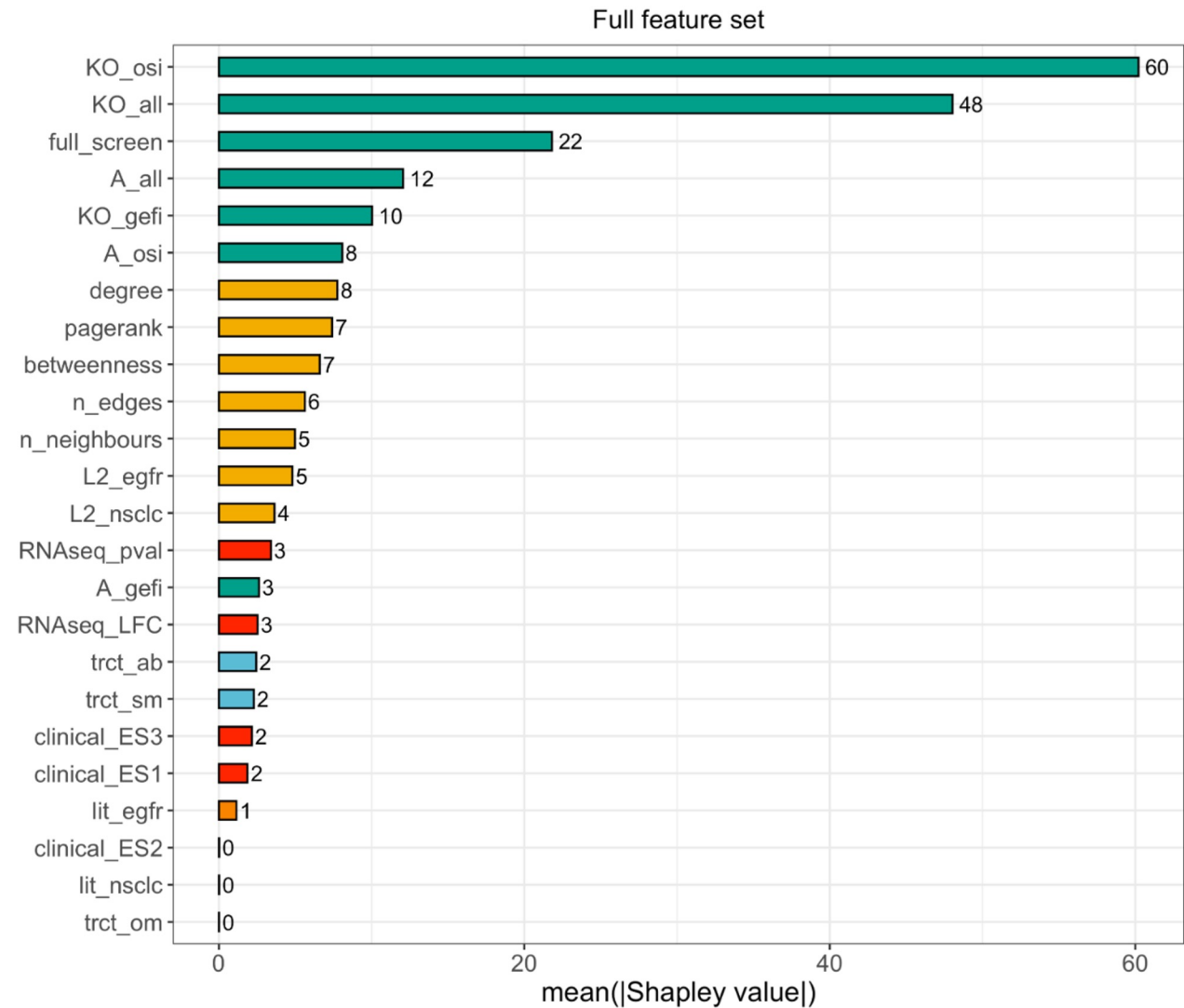
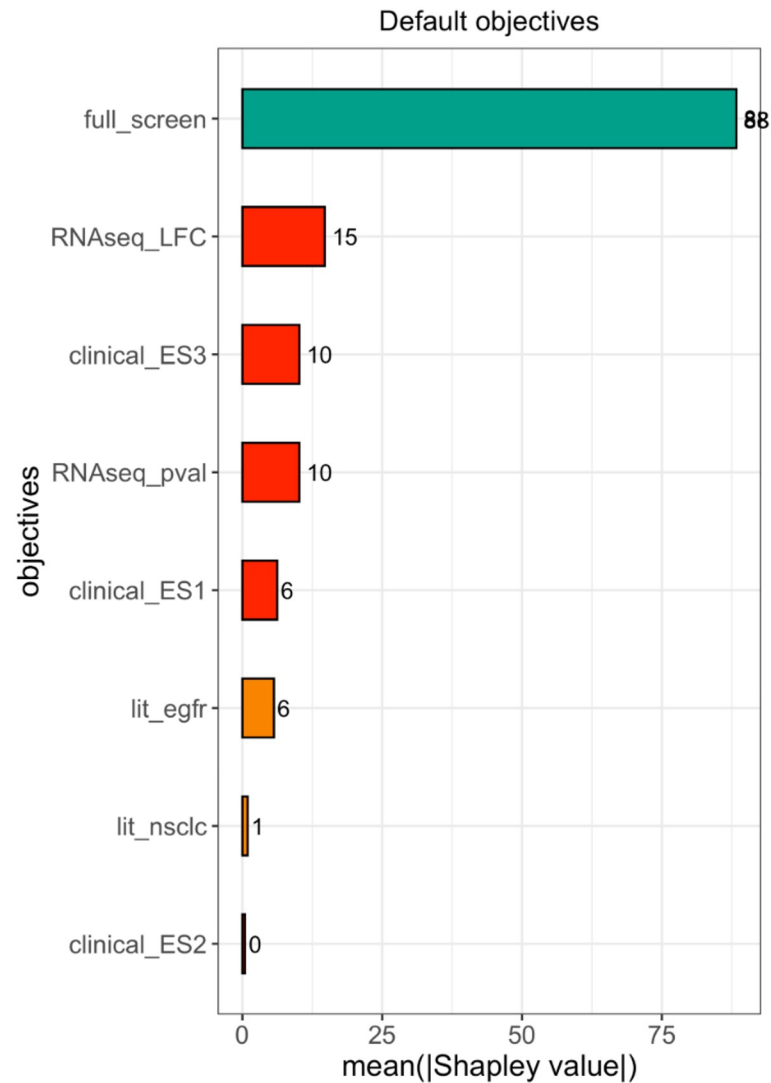
Below the table, it says 'Showing 1 to 10 of 42 entries' and has a 'Download Top results' button.

On the right side of the interface, there are two heatmaps. The top one is titled 'heatmap controls' and has sliders for 'min cluster size' (set to 3) and 'min number of papers with gene cluster' (set to 50). Below it is a dropdown menu 'select a gene, only genes found in NLP clusters are shown' with 'ENSG0000017' selected. The bottom heatmap is titled 'heatmap showing multi-term gene co-occurrence in cancer resistance context' and shows a complex grid of colored squares representing gene co-occurrence, with a list of genes on the right including KRAS, BRAF, EGFR, PIK3CA, etc.

Imperfect validation



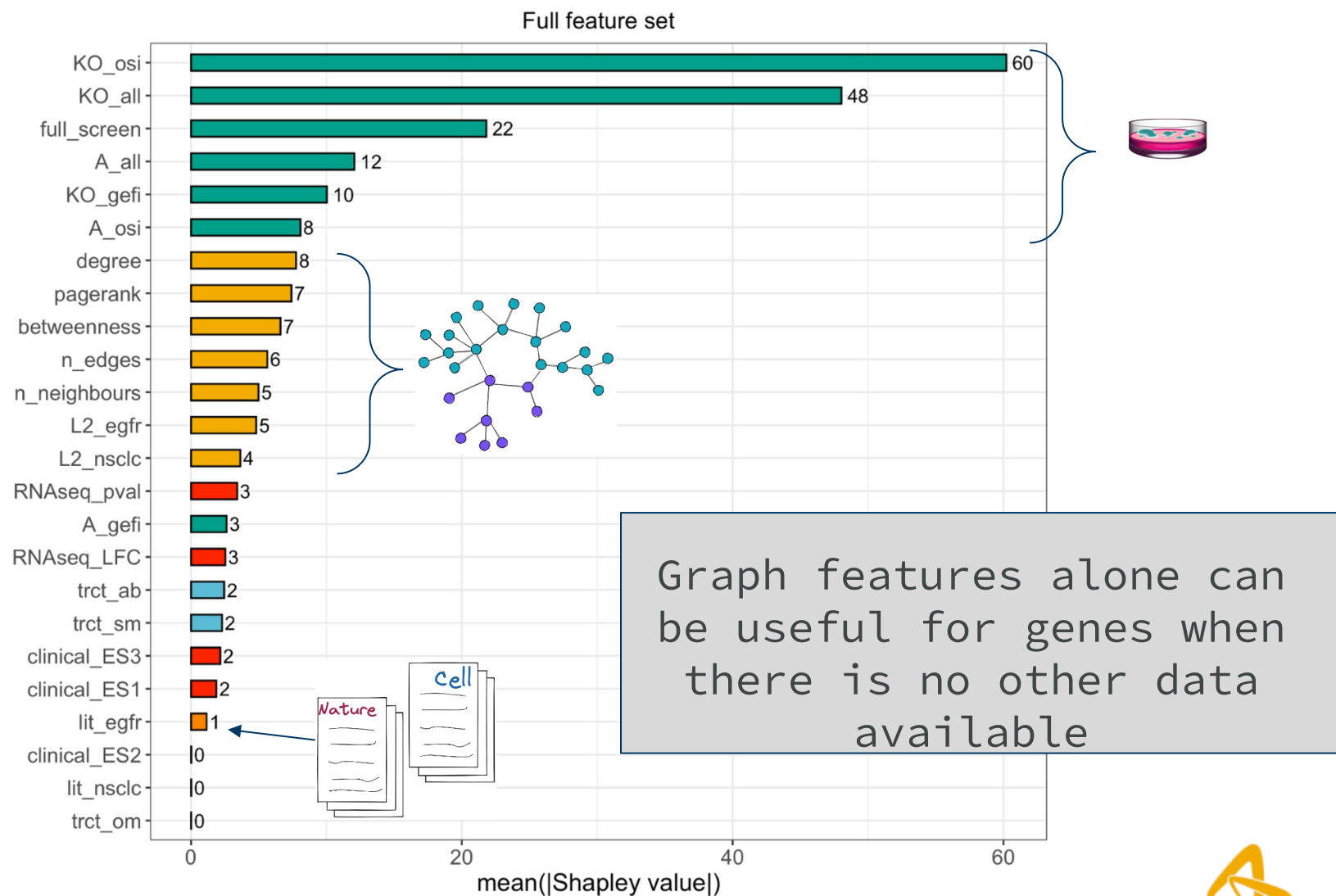
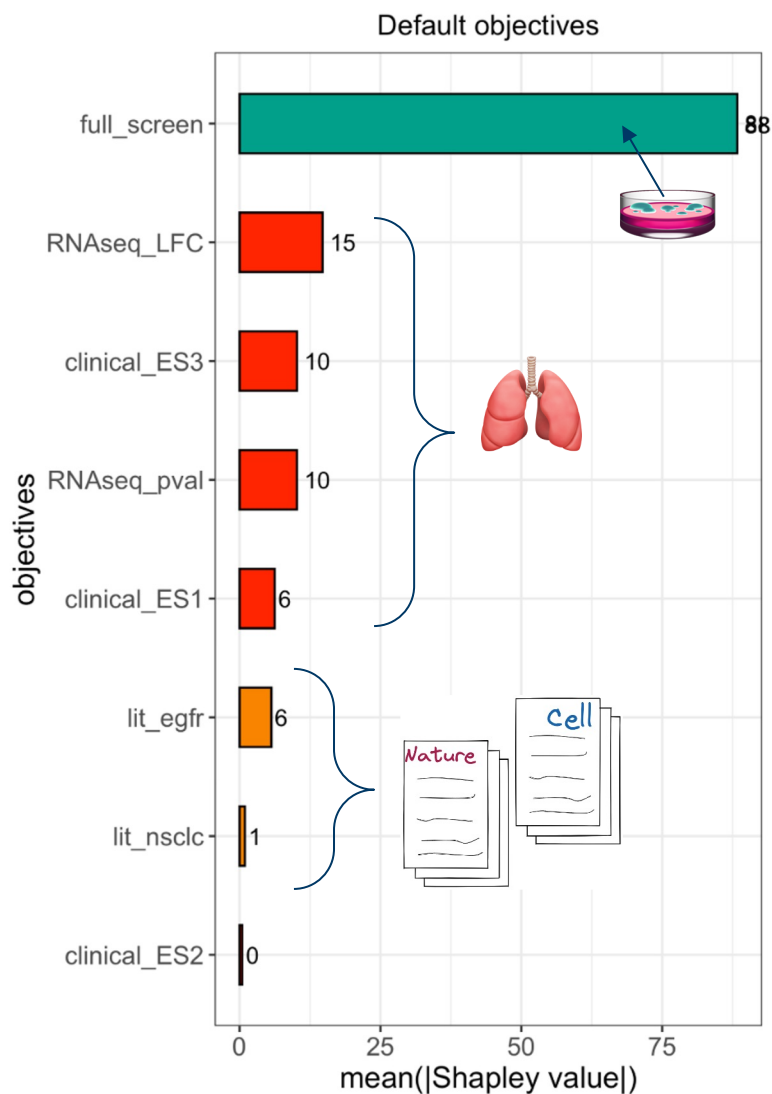
Model domain scientist as a black box classifier



Category ■ Clinical/Preclinical ■ CRISPR ■ Graph ■ NLP ■ Tractability



Graph-derived features follow clinical in unbiased setting



Annotation by the experts

WWTR1 WW domain containing transcription regulator 1
ENSG00000018408

#Publications of this hit mentioned within the context of 'resistance' and 'EGFR': 0
#Publications of this hit mentioned within the context of 'resistance' and 'NSCLC': 0

for additional evidence behind the gene recommendation please see [skywalkB](#)

Known resistance marker 1

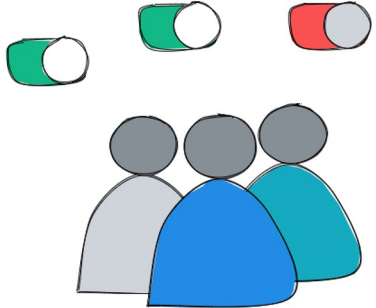
Novel, but credible hit 2

Novel, not credible hit 3

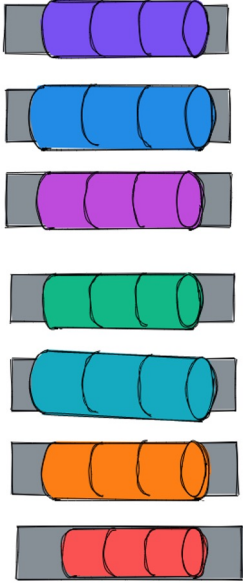
Not novel, not credible hit 4

please include any additional details about ongoing experiments for this marker, or if this has been discussed at (pre)TSID.

TASK_NUM: 1 TOTAL_TASKS_NUM: 42



Gene list



Most of recommendations are 'novel & credible'

WWTR1 WW domain containing transcription regulator 1
ENSG0000018408

#Publications of this hit mentioned within the context of 'resistance' and 'EGFR': 0
 #Publications of this hit mentioned within the context of 'resistance' and 'NSCLC': 0

for additional evidence behind the gene recommendation please see [skywalkB](#)

Known resistance marker 1

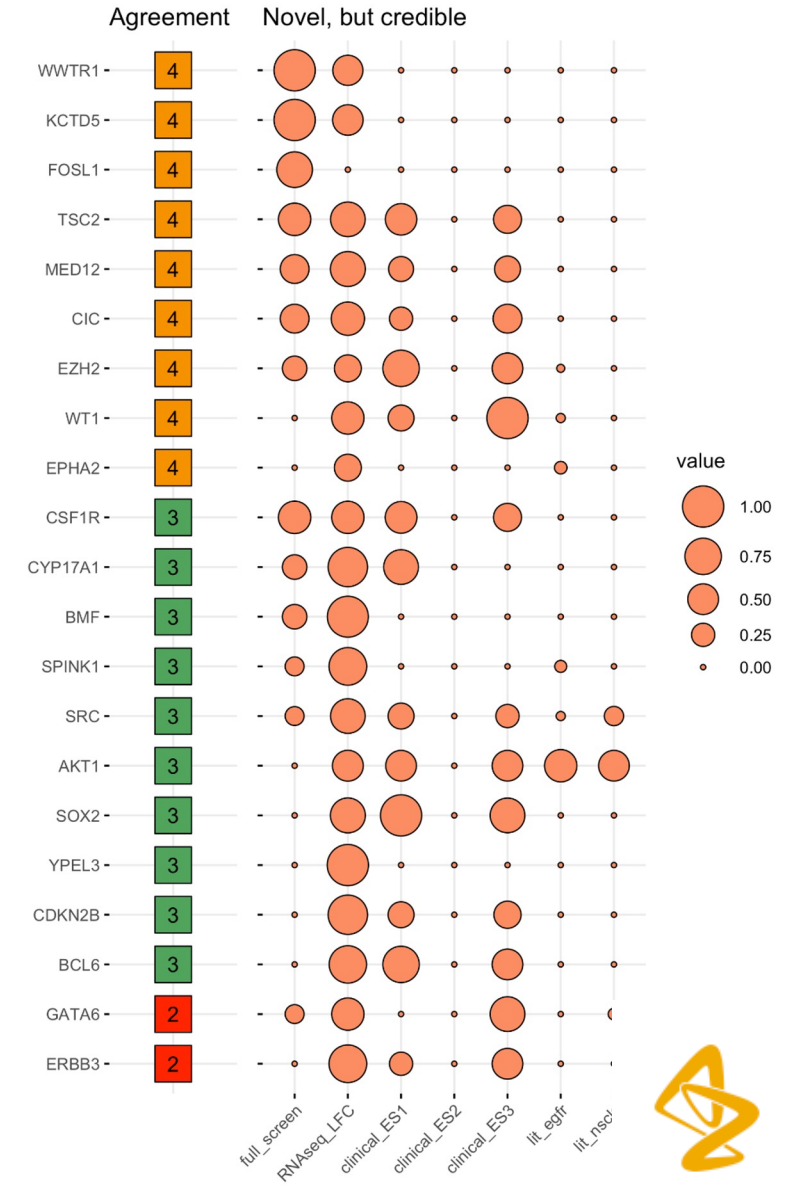
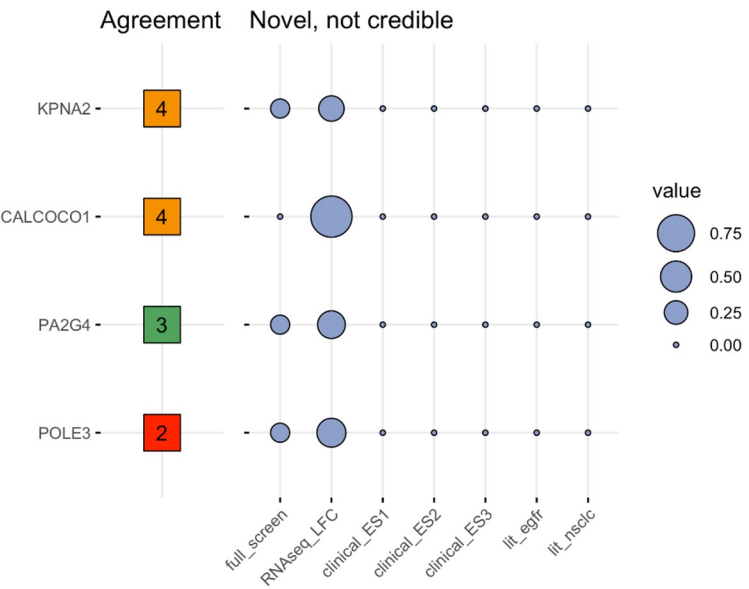
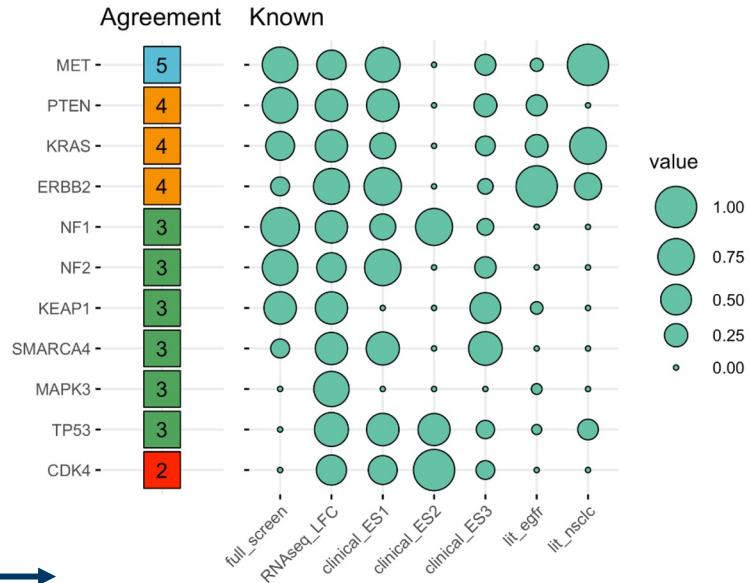
Novel, but credible hit 2

Novel, not credible hit 3

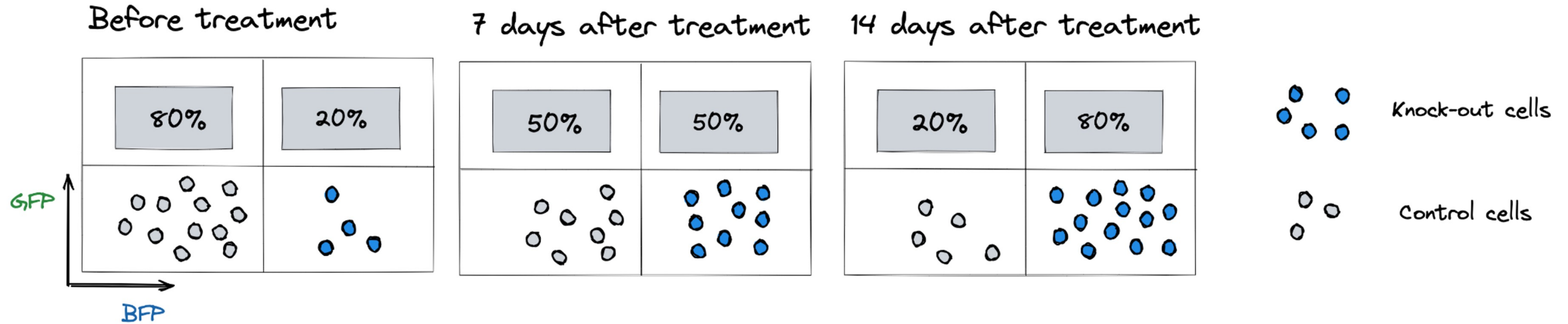
Not novel, not credible hit 4

please include any additional details about ongoing experiments for this marker, or if this has been discussed at (pre)TSID.

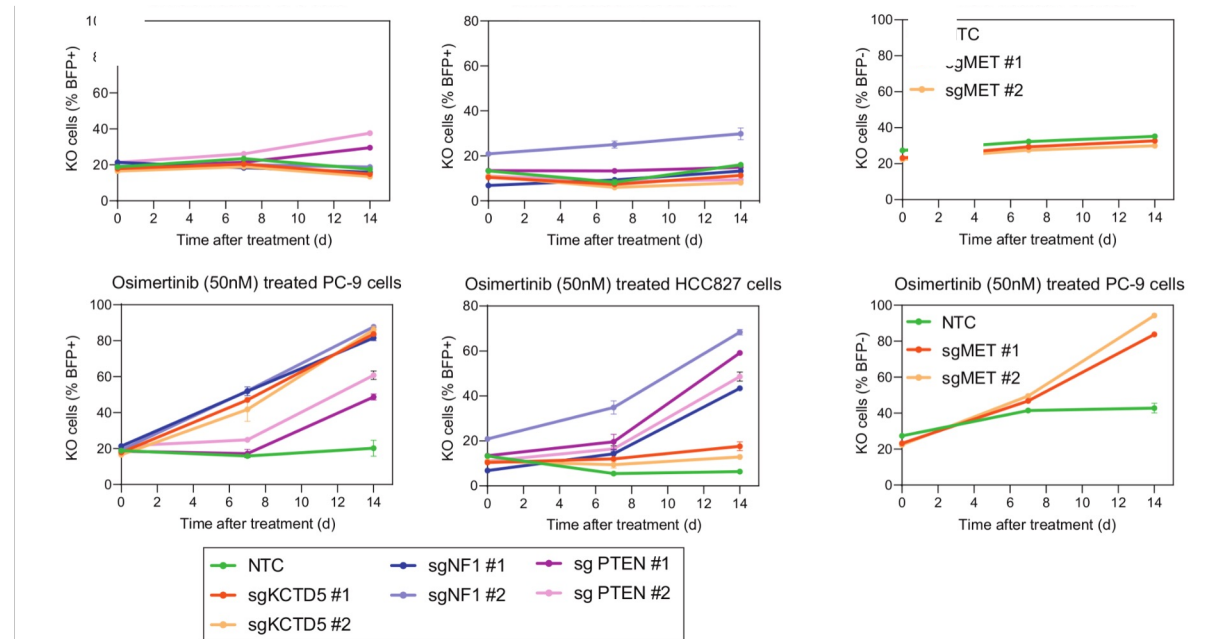
TASK_NUM: 1 TOTAL_TASKS_NUM: 42



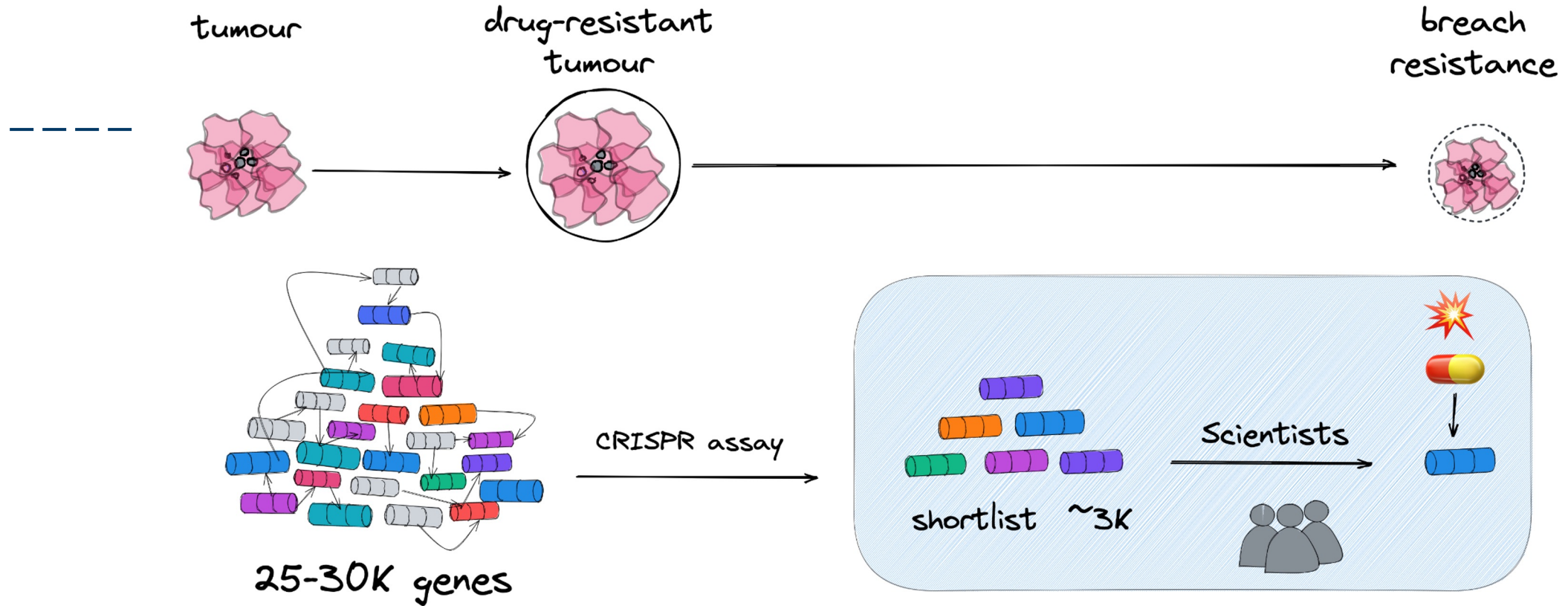
Experimental validation *in vitro*



- confirmed involvement of 6 recommended genes in drug resistance
- next: test the remaining genes



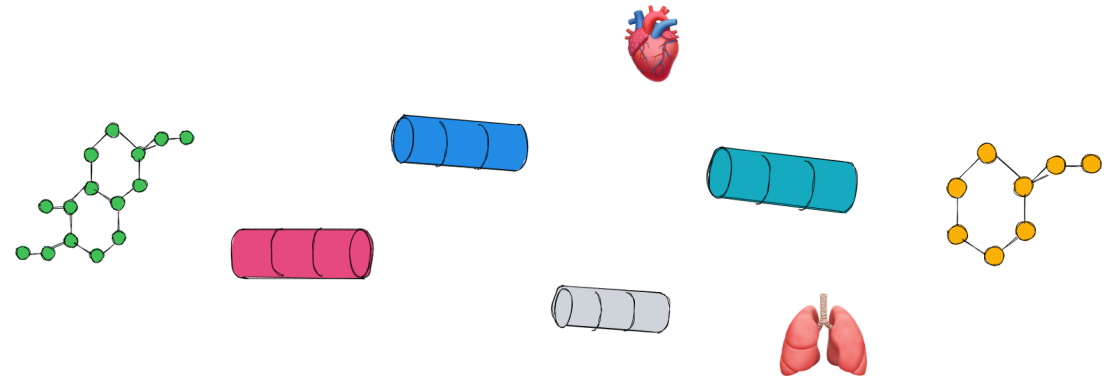
Imperfect, yet already useful recommendation system



- 🐌 -> 🚗 re-rank lists in seconds, not months
- ⚙️ automated feature generation
- ♻️ approach can be re-used in related problems
- 👍 now a standard solution for CRISPR screens



Take home message



- Drug discovery is an exciting field for recommender systems
- Relatively simple recommenders can have a lot of impact
- Need for recommenders that can operate in unsupervised or weakly supervised settings
- There are a number of challenges



Acknowledgements

Early Computational Oncology @ AZ

Krishna Bulusu

Ben Sidders

Daniel Barrell

Miika Ahdesmäki

Jonathan R. Dry

R&D IT @ AZ

Vladimir Poroshin

Michaël Ughetto

Eliseo Papa

Bioscience, Oncology R&D @ AZ

Matthias Pfeifer

Ultan McDermott



We are hiring 🐱💻!

